

# Phylogenetic closure operations, and homoplasy-free evolution

Mike Steel

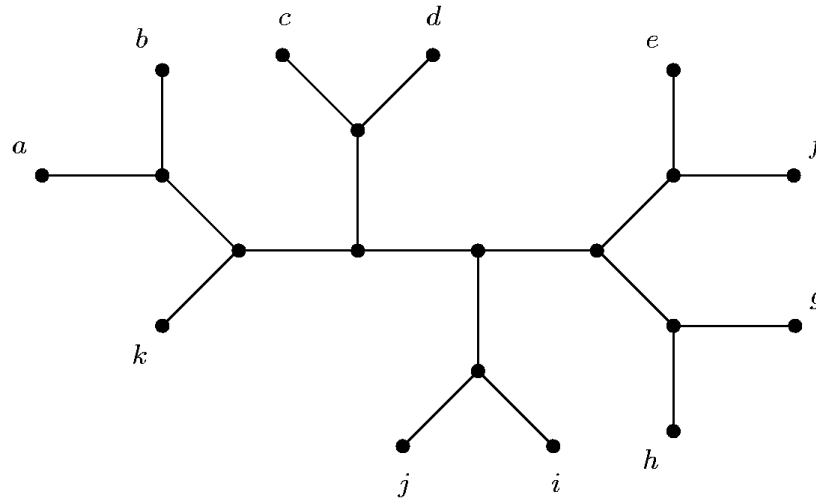
Biomathematics Research Centre  
University of Canterbury, Christchurch, New Zealand



*Joint work with* Tobias Dezulian,  
Center for Bioinformatics Tübingen,  
University of Tübingen, Germany

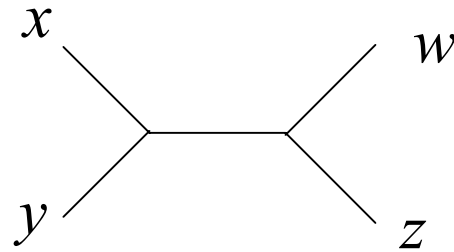
# Phylogenetic trees

- [Definition] A **phylogenetic X-tree** is a tree  $T=(V,E)$  with a set  $X$  of labelled leaves, and all other vertices unlabelled and of degree  $\geq 3$ .
- If all non-leaf vertices have degree 3 then  $T$  is **binary**

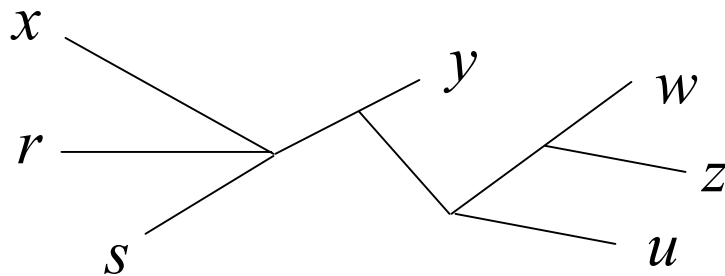


# Quartet trees

- A **quartet tree** is a binary phylogenetic tree on 4 leaves (say,  $x, y, w, z$ ) written  $xy|wz$ .



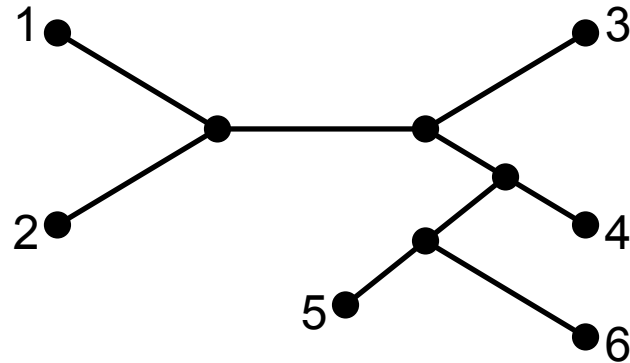
- A phylogenetic X-tree **displays**  $xy|wz$  if there is an edge in  $T$  whose deletion separates  $\{x, y\}$  from  $\{w, z\}$



# Compatibility

A set  $Q$  of quartets is compatible if there is a phylogenetic  $X$ -tree  $T$  that **displays** each quartet of  $Q$

- **Example:**  $Q = \{12|34, 13|45, 24|56\}$



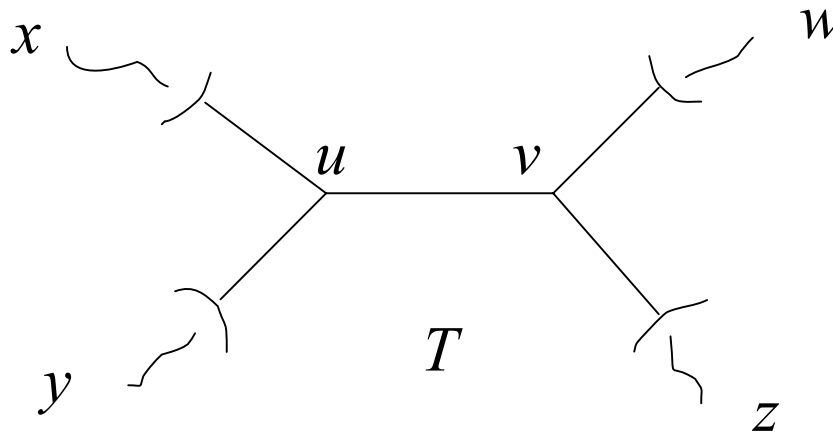
# Defining sets

If  $T$  is the only phylogenetic  $X$ -tree that displays  $Q$  (and  $X = L(Q)$ ) then we say  $Q$  **defines**  $T$ .

- Let  $Q(T)$  be the set of **all** quartets displayed by (any)  $T$ .  
If  $T$  is binary, then  $Q(T)$  defines  $T$ .

## Definition:

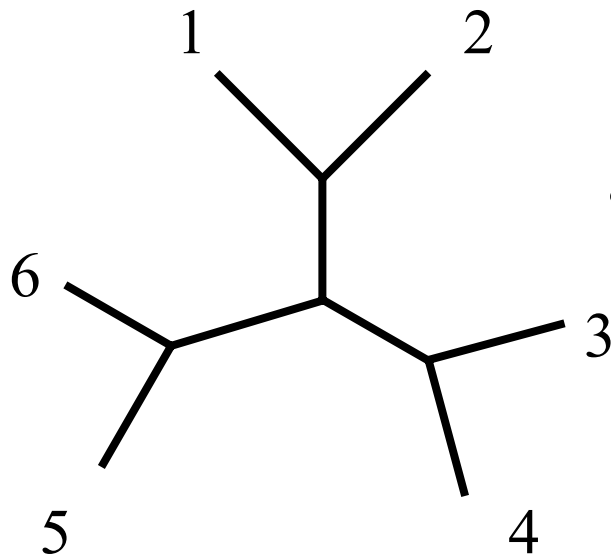
- For a binary phylogenetic tree  $T$ , a collection  $Q$  of induced quartet trees *distinguishes* an interior edge  $\{u,v\}$  of  $T$  if there exists a quartet  $xy|wz$  in  $Q$  that looks like this:



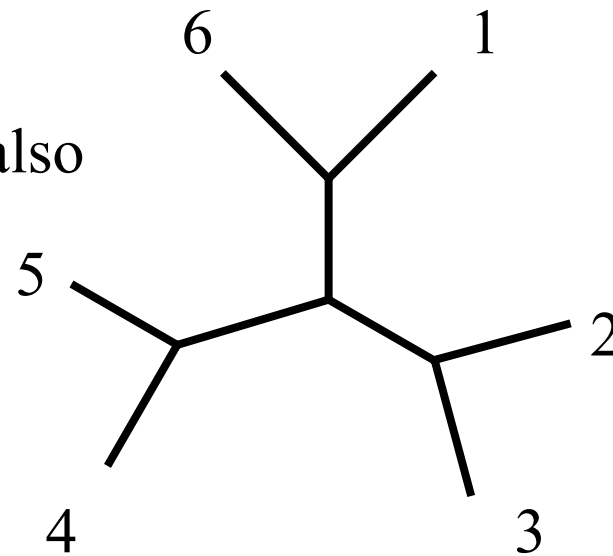
**Observation:** If  $Q$  defines  $T$  then  $T$  is binary and  $Q$  distinguishes every interior edge of  $T$  (so  $|Q| \geq n-3$ ).

# Warning:

$Q = \{12|45, 56|23, 34|16\}$  distinguishes each interior edge of the tree:



and also



!

## Sufficient condition for $Q$ to define $T$ :

- Suppose  $Q$  is compatible and distinguishes every interior edge of a binary phylogenetic  $X$ -tree  $T$ .

**Proposition:** If there is an element of  $X$  that is a leaf of every tree in  $Q$  then  $Q$  defines  $T$ .

**Corollary:**

There are subsets of  $Q(T)$  that define  $T$  of size  $|X|-3$ .

# Character data

■ Type	States	Transitions
■ Morphology	$W(\text{ings}), \neg W, -W$	$\neg W \rightarrow W \rightarrow -W$
■ Sequences	A, C, G, T	$x \leftrightarrow y$
■ Gene order	$g_1 g_2 g_3 g_4 g_5 g_6 g_7 \dots$ $\dots g \dots$	$g_1 g_2 \boxed{g_5 g_4 g_3} g_6 g_7 \dots$ $\dots \rightarrow \dots g \dots \rightarrow \dots ? \dots$
■ SINEs		

# Definitions:

- [Character] A character is any function

$$f : X \rightarrow S$$

- [Convexity] Given a character  $f : X \rightarrow S$

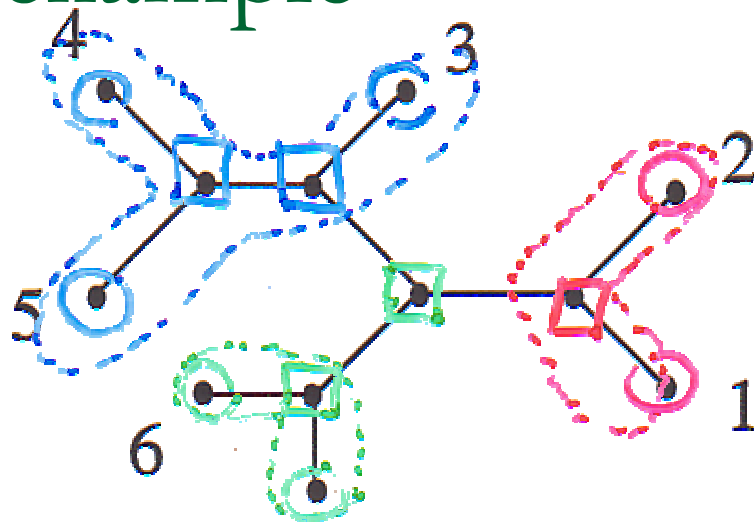
and a phylogenetic  $X$ -tree  $T=(V,E)$ , we say  $f$  is **convex on  $T$**

if  $f$  extends to  $f':V \rightarrow S$

so that  $f'|_X = f$

and  $\{v \in V : f'(v) = s\}$  is connected for all  $s$  in  $S$ .

# Convexity: example

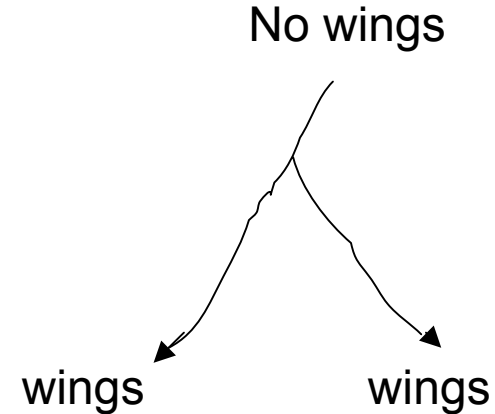
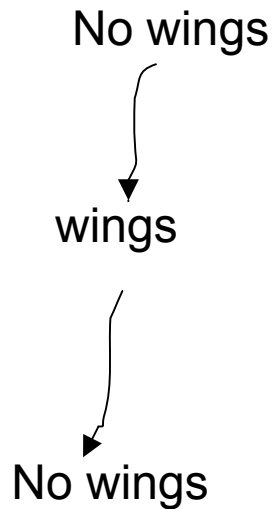


$x$	1	2	3	4	5	6	7
$f(x)$	●	●	●	●	●	●	●

# Biological significance of convexity

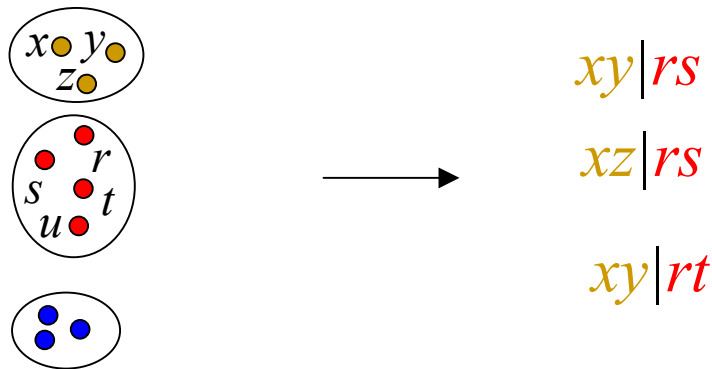


- Lemma: A character  $\chi$  is convex on a phylogenetic tree  $T$  if and only if  $\chi$  could have evolved on  $T$  (from any root vertex) without any **reversals** or **convergent evolution**.



# Equivalence of character and quartet compatibility

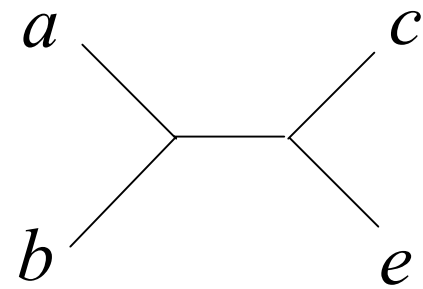
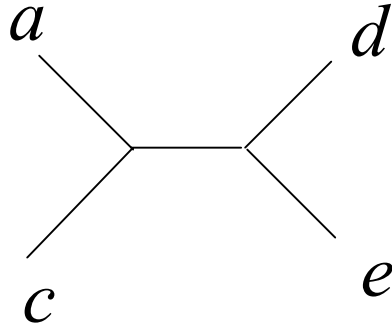
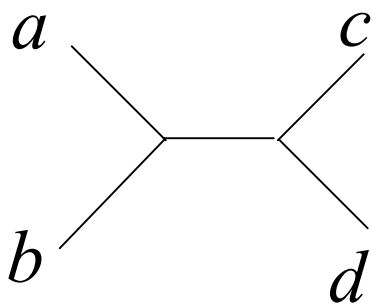
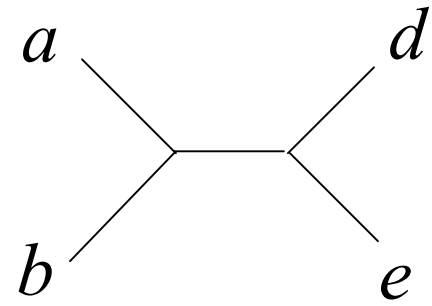
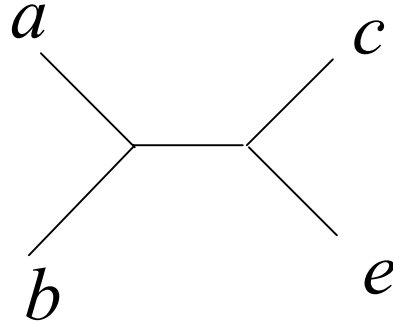
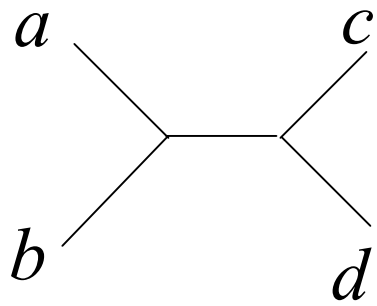
$$C \rightarrow Q(C)$$



**Lemma:** Each character in  $C$  is convex on  $T$  if and only if  $T$  displays all the quartets in  $Q(C)$ .

[ $C$  is “compatible”,  $C$  “defines”  $T$  iff  $Q(C)$  does]

# New quartet trees from old ones



# Dyadic rules for quartet trees

(Coloniuss and Schulze; Dekker)

(**Q1**):  $\{ab|cd, ab|ce\} \vdash ab|de$

(**Q2**):  $\{ab|cd, ac|de\} \vdash ab|ce, ab|de, bc|de.$

Any phylogenetic  $X$ -tree that displays the quartet trees on the left of (**Q1**) or (**Q2**) also displays the corresponding quartet tree(s) on the right.

# Dyadic quartet closure

$$\mathcal{Q} = \mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \cdots \subseteq \mathcal{Q}_m = \text{qcl}_\theta(\mathcal{Q})$$

where  $\mathcal{Q}_{i+1}$  consists of  $\mathcal{Q}_i$  together with all additional quartets that can be obtained from a pair of quartets in  $\mathcal{Q}_i$  by applying the rule(s) allowed by  $\theta$ .

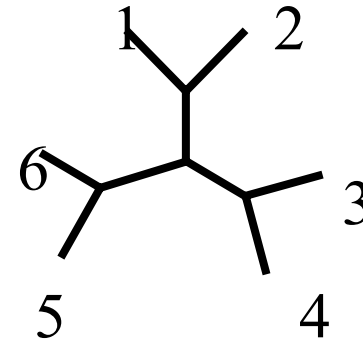
For  $\theta \subseteq \{1, 2\}$ , let the dyadic quartet closure under rule  $\theta$ ,  $\text{qcl}_\theta(\mathcal{Q})$ , denote the minimal set of quartet trees that contains  $\mathcal{Q}$  and is closed under rule **(Qi)** for each  $i \in \theta$ .

We denote these closures with:  $\text{qcl}_1(\mathcal{Q})$ ,  $\text{qcl}_2(\mathcal{Q})$ ,  $\text{qcl}_{1,2}(\mathcal{Q})$ .

# Example 1: $qcl_2$

**Definition:** If  $Q$  distinguishes every interior edge of a binary phylogenetic tree  $T$  and we can order  $Q$  so that each quartet tree in the ordering introduces precisely one new leaf label, we say  $Q$  has a **tight ordering** for  $T$ .

Example:  $\{12|35, 13|56, 15|34\}$ .



**Proposition:**

If  $Q$  has a tight ordering for  $T$ , then  $qcl_2(Q) = Q(T)$

In particular  $Q$  defines  $T$ .

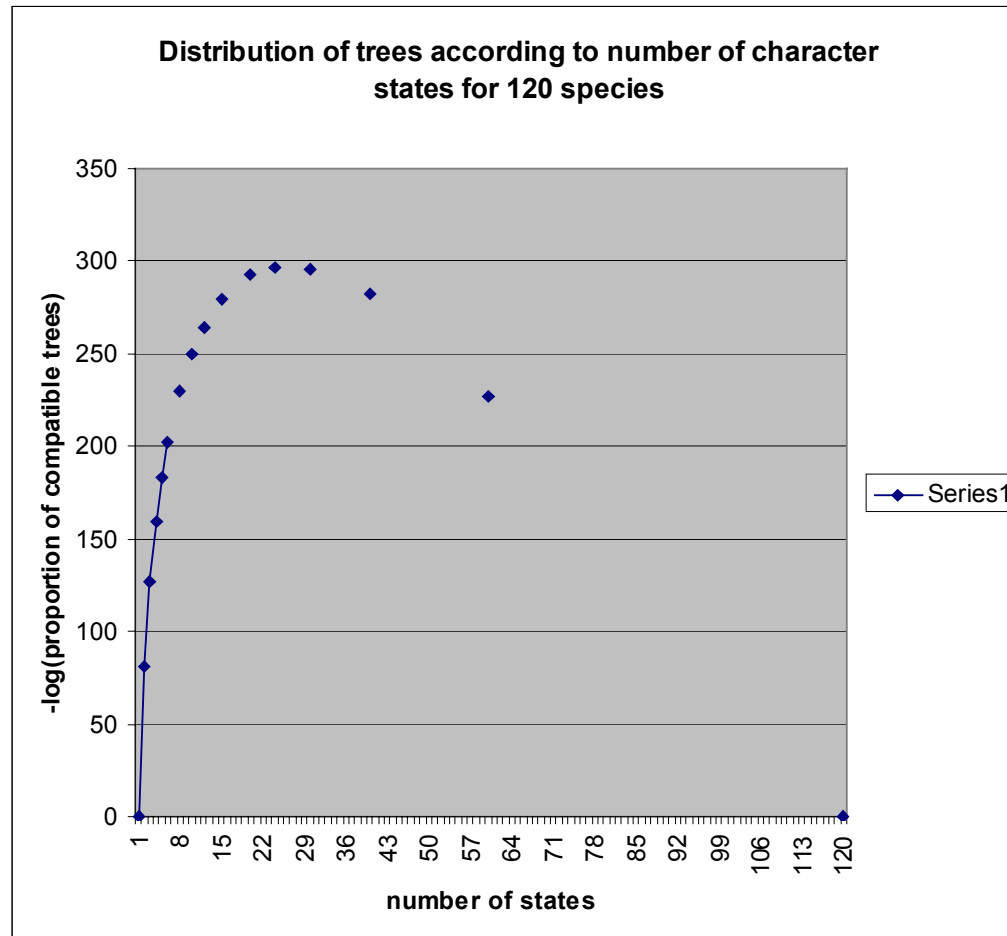
**Application:** How many characters are needed to define a binary phylogenetic  $X$ -tree?

- For binary characters we need  $n-3$  ( $n=|X|$ ).
- For  $r$ -state characters ( $r$  fixed) we need at least  
 $(n-3)/(r-1)$
- What if  $r$  is not fixed?

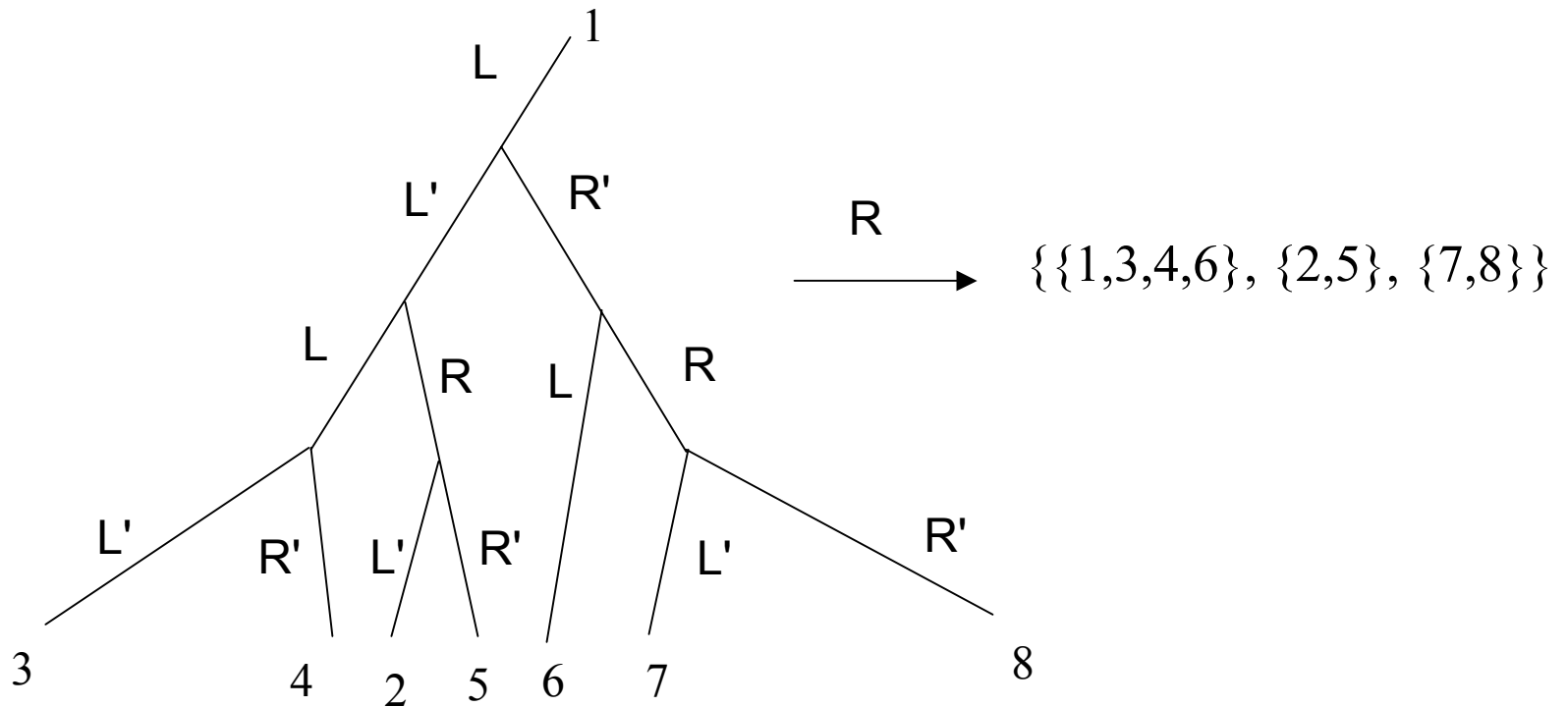
(it is not useful to make  $r$  too large!)

$$I(\chi) := -\log(\Pr[\chi \text{ is convex on random } T])$$

= explicit formula (sums of logs of odd integers), Carter *et al.* (1990); Erdős & Székely (1993).



# Edge-colouring a tree by $Z_2 \times Z_2$



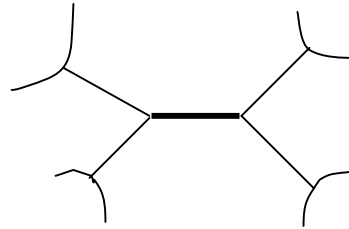
**Theorem** (Huber, Moulton, Steel 2003)

$Q(C)$  contains a subset with a tight ordering for  $T$ .

Thus for any tree there is a set of just FOUR characters that defines  $T$ .

## Application 2: “Short” quartets

■  $Q_{\text{short}}(T)$



■ **Theorem** (Erdős et al. 1997)

$Q_{\text{short}}(T)$  contains a subset that has a tight ordering for  $T$   
(and so  $\text{qcl}_2(Q_{\text{short}}(T)) = Q(T)$ ).

■ The number of characters required to reconstruct (wp  $> 1 - \varepsilon$ ) a binary phylogenetic tree with  $n$  leaves from binary characters generated under a finite Markov process is (for almost all trees) at most

$$k \geq \frac{c_\varepsilon (\log(n))^{d(p)}}{a^2}$$

# A further application involving $qcl_2$ :

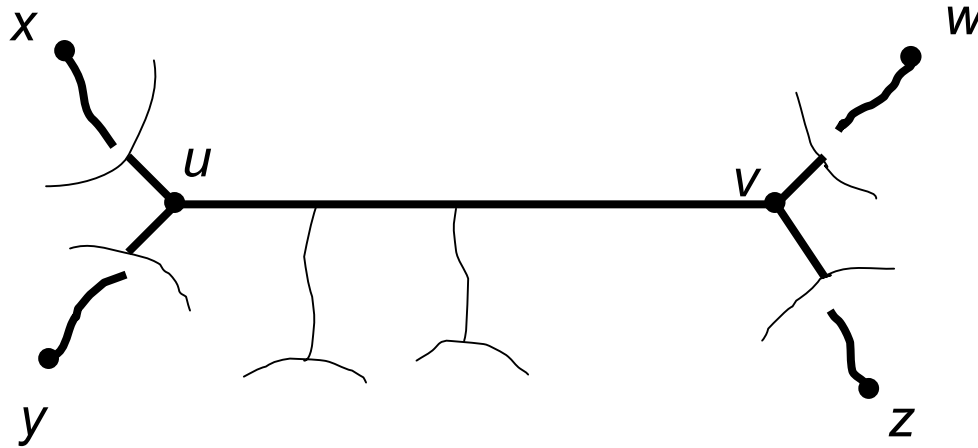
We say  $Q$  is **excess-free** if  $|L(Q)| - 3 - |Q| = 0$ .

- **Proposition:** Suppose a subset  $Q$  of  $Q(T)$  contains an excess-free subset  $Q_0$  that defines  $T$ . Then  $qcl_2(Q) = Q(T)$ .
- **Why?** Let us say a set  $Q$  of quartet trees is “good” if (i)  $Q$  defines a phylogenetic tree, and (ii)  $exc(Q) = 0$ .

**Theorem** [Bocker, Dress 1999] Any good set of ( $\geq 2$ ) quartets is the disjoint union of precisely two good sets.

## Example 2: $qcl_1$

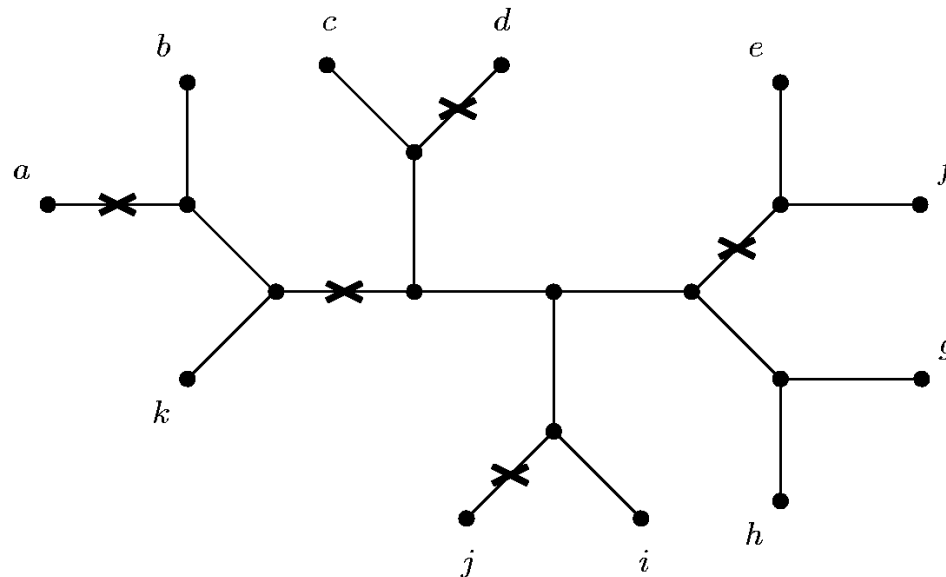
- **Definition:** For a binary phylogenetic tree  $T$ , a collection  $Q$  of displayed quartet trees is a *generous cover* for  $T$  if for all pairs  $u, v$  of interior vertices of  $T$ , we have a quartet  $xy|wz$  in  $Q$  that looks like this:



**Theorem** (Dezulian + S, 2003): If  $Q$  is a generous cover for  $T$ , then  $qcl_1(T) = Q(T)$ . Thus  $Q$  defines  $T$ .

# Application: the random cluster model

Random process on a phylogenetic tree  $\mathcal{T}$ . Independently cutting edges with probability  $p(e)$  generates, by connectivity, random characters on  $\mathcal{T}$ .



Cutting the marked edges yields the character  $\{a|bk|cghi|d|ef|j\}$ .

## Theorem (Mossel and S, 2003)

- For random cluster model, if  $0 < a \leq p(e) \leq p < 0.5$ , every binary phylogenetic tree with  $n$  leaves can be reconstructed with probability at least  $1 - \varepsilon$  from  $k$  characters if

$$k \geq \frac{c_{p,\varepsilon} \log(n)}{a}$$

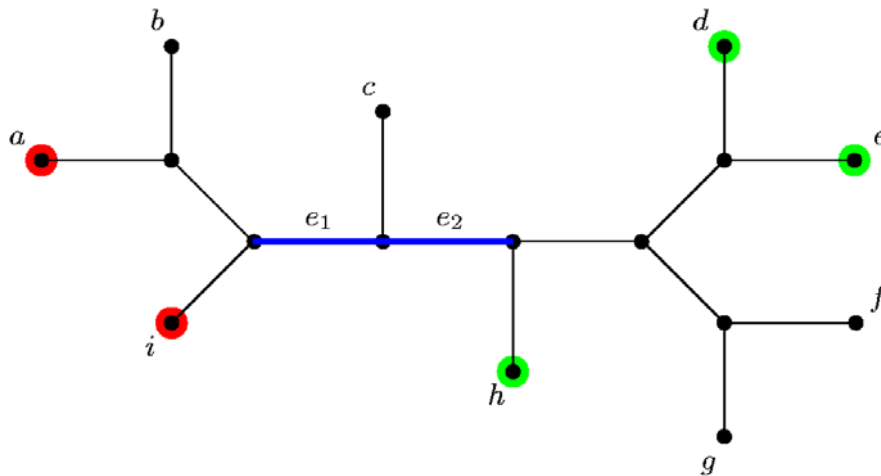
- A fast (polynomial-time) algorithm to reconstruct  $T$  from the characters.
- Proof uses generous cover result
- Lower bound:  $\log(n)$  needed (not trivial) and polynomial ( $n$ ) if  $p > 0.5$ .
- *cf.* finite-state

$$k \geq \frac{c_\varepsilon (\log(n))^{d(p)}}{a^2}$$

# Application 3: $qcl_1, qcl_2, qcl_{1,2}$

**[Definition]** A **partial X-split**  $A|B$  is a partition of a subset into two non-empty sets,  $A, B$ .  $A|B$  is **displayed** by  $T$  if we can remove an edge from  $T$  to separate  $A$  from  $B$ .

**Example:**  $\{a, i\} | \{d, e, h\}$  is displayed by  $T$ .



## Christopher Meacham's dyadic rules for splits (1983)

**(M1)**: If  $A_1 \cap A_2 \neq \emptyset$  and  $B_1 \cap B_2 \neq \emptyset$  then

$\{A_1|B_1, A_2|B_2\} \vdash A_1 \cap A_2 | B_1 \cup B_2, A_1 \cup A_2 | B_1 \cap B_2.$

**(M2)**: If  $A_1 \cap A_2 \neq \emptyset$  and  $B_1 \cap B_2 \neq \emptyset$  and  $A_1 \cap B_2 \neq \emptyset$  then

$\{A_1|B_1, A_2|B_2\} \vdash A_2 | B_1 \cup B_2, A_1 \cup A_2 | B_1.$

Any phylogenetic  $X$ -tree that displays the partial  $X$ -splits on the left of **(M1)** or **(M2)** also displays the corresponding partial  $X$ -splits on the right.

## Dyadic split closure

$$\Sigma = \Sigma_1 \subseteq \Sigma_2 \subseteq \cdots \subseteq \Sigma_m = \text{spcl}_\theta(\Sigma)$$

where  $\Sigma_{i+1}$  consists of  $\Sigma_i$  together with all additional splits that can be obtained from a pair of splits in  $\Sigma_i$  by applying the rule(s) allowed by  $\theta$ .

For  $\theta \subseteq \{1, 2\}$ , let the dyadic split closure under rule  $\theta$ ,  $\text{spcl}_\theta(\Sigma)$ , denote the minimal set of splits that contains  $\Sigma$  and is closed under rule **(Mi)** for each  $i \in \theta$ .

We denote these closures with:  $\text{spcl}_1(\Sigma)$ ,  $\text{spcl}_2(\Sigma)$ ,  $\text{spcl}_{1,2}(\Sigma)$ .

## the (almost) happy marriage

$$\begin{array}{ccc} \Sigma & \xrightarrow{\mathcal{Q}} & \mathcal{Q}(\Sigma) \\ \text{spcl}_\theta \downarrow & & \downarrow \text{qcl}_\theta \\ \text{spcl}_\theta(\Sigma) & \xrightarrow{\mathcal{Q}} & (*) \end{array}$$

?

## the (almost) happy marriage

$$\begin{array}{ccc} \Sigma & \xrightarrow{\mathcal{Q}} & \mathcal{Q}(\Sigma) \\ \text{spcl}_\theta \downarrow & & \downarrow \text{qcl}_\theta \\ \text{spcl}_\theta(\Sigma) & \xrightarrow{\mathcal{Q}} & (*) \end{array}$$

**Theorem 2.1.** *Let  $\Sigma$  be a collection of partial  $X$ -splits. Then,*

$$\text{qcl}_\theta(\mathcal{Q}(\Sigma)) = \mathcal{Q}(\text{spcl}_\theta(\Sigma))$$

*for  $\theta = \{1\}$  and  $\theta = \{1, 2\}$ . For  $\theta = \{2\}$  we have*

$$\text{qcl}_\theta(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\text{spcl}_\theta(\Sigma))$$

*-and containment can be strict.*

---

## Further details

- A phase transition for a random cluster model on phylogenetic trees. E. Mossel and M. Steel, *Mathematical Biosciences*, 187 (2004), 189-203.
- Phylogenetic closure operations, and homoplasy-free evolution, T. Dezulian and M. Steel *Proceedings of the International Federation of Classification Societies*, Chicago, 2004.